An experimental comparison between NMF and LDA for active cross-situational object-word learning

Yuxin Chen, Jean-Baptiste Bordes, David Filliat

An experimental comparison between NMF and LDA for active cross-situational object-word learning

Yuxin Chen* Jean-Baptiste Bordes*† David Filliat*
*U2IS, ENSTA ParisTech, Inria FLOWERS team, Université Paris Saclay, Palaiseau, France
†Ecole Polytechnique, Palaiseau, France

***Abstract*—Humans can learn word-object associations from ambiguous data using cross-situational learning and have been shown to be more efficient when actively choosing the learning sample order. Implementing such a capacity in robots has been performed using several models, among which are the latent-topic learning models based on Non-Negative Matrix Factorization and Latent Dirichlet Allocation. We compare these approaches on the same data in a batch and in an incremental learning scenario to analyze their strength and weaknesses and furthermore show that they can be the basis for efficient active learning strategies. The proposed modeling deals with both the referential ambiguity and the noisy linguistic descriptions and is grounding meanings of object's modal features (color and shape) and not only the object identity. The resulting active learning strategy is briefly discussed in comparison with active cross-situational learning of object names performed by humans.**

I. Introduction

Learning new words describing objects and their meanings during direct interaction between a robot and a human is a challenging task. This problem, related to the symbol grounding problem [1], faces several sources of ambiguities. *Linguistic ambiguity* exists in the words to be learned as the human pronounces complex sentences where not all the words are relevant for describing an object (such as pronouns or verbs). *Referential ambiguity* is present in the described object when the robot is facing a complex scene where multiple objects appear in its field of view (Figure 1).

When learning word-object associations, human use several strategies to reduce these ambiguities. The *linguistic ambiguity* may be reduced by taking advantage of the grammar that will highlight the relevant nouns and adjectives in a sentence [2]. The *referential ambiguity* may be reduced by the use of joint attention that makes both teacher and learner focus on the same object [3]. Nevertheless, it has been shown that humans and infants, as young as 12 months old, can learn in ambiguous situations by relying on cross-situational learning [4], i.e, by analyzing the common factors between several ambiguous situations displaying various objects and associated words. In this paper, we therefore focus on the problem of learning from ambiguous data, rather than studying the techniques that could reduce this ambiguity, while in complete application scenario, both approaches should obviously be applied.

Several models of cross-situational learning have been proposed (e.g., [5], [6], [7]). In particular, this problem can be treated by latent topic discovery approaches [8] that will find the underlying element (the object or feature) that will generate both the visual perception and the associated word. Among these approaches, we focus in this paper on comparing Non Negative Matrix Factorization (NMF) [9] and Latent Dirichlet Allocation (LDA) [10] that have been used to model cross-situational learning in different setups [11], [12].

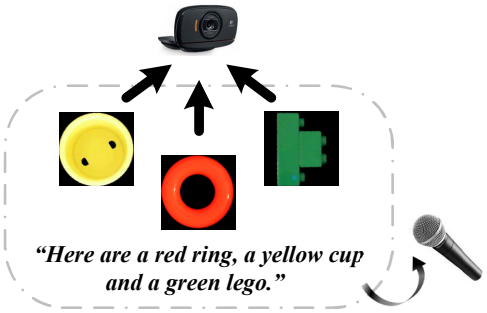

Fig. 1. Example of an ambiguous teaching situation presenting referential and linguistic ambiguities.

Another strategy used by humans to improve learning is active learning, where learners choose the learning samples so as to improve learning speed. Implementation of active learning may rely on intrinsic motivations and computational models with this capacity have been proposed in developmental robotics [13] and studied for the problem of language learning [14]. However, the definition of the intrinsic motivation often depends on the task and the particular learning algorithm. In our case, it is therefore not clear if both NMF and LDA would be well suited to implement active learning.

Our main contribution is a detailed comparison of the performances of NMF and LDA for the task of learning the meaning of nouns and adjectives describing objects in ambiguous setups. We focus on their performance evolution while incrementally learning from a limited set of examples and explore whether they are well suited to define a value function for active learning. Unlike most models of cross-situational learning [4], [5], [6], [7], [12], [15], we moreover focus on the learning of object's descriptive features instead of only its identity and deal with the linguistic ambiguity besides the classical referential ambiguity in experimental settings. Tackling both these ambiguities, our model lays the basis for

interactive learning of concept for developmental robots.

In the remainder of the paper, we review the related work in the next section, then present our application of NMF and LDA to cross-situational and active learning before presenting a quantitative comparison of these approaches on a dataset of objects described by human teachers.

II. Related work

Several models of cross-situational learning have been proposed using different techniques such as hypothesis testing and associative learning [7], Expectation-Maximisation [6] or measures of co-occurences and mutual information [5]. In this paper, we focus on latent topic discovery approaches [8] that are well suited to this problem. The main idea is to find a limited number of hidden topics that explain the data. In our case, a *topic* would be an object or a colour that would generate both its visual perception and its associated name. This definition is closely related to *concept*, ie. mental representation of patterns in a flow of multimodal perception (see [16] for a more in-depth discussion).

Among the existing topic discovery algorithms, two have been used for cross-situational learning: Non Negative Matrix Factorization in [16], [11] and Latent Dirichlet Allocation in [12], [17]. However, they are applied in different settings and it is not clear which one is better-suited for this task. We propose here a direct comparison to highlight their strength and weaknesses, particularly in the case where the number of training samples is limited as when they are acquired through direct interaction between a human and a robot.

Active learning by the use of intrinsic motivations has been proposed as a way to control the complexity of learning situations so as to improve learning speed and coverage of the learnable space [13]. It has been applied to many sensorymotor skills, and has also been argued as one of the important bias for learning language [18]. Indeed, active learning has been shown to be a factor strongly improving learning quality in cross-situational learning for humans [15]. Several specific strategies have also been proposed for the application of active learning to language learning computational models in [14] by controlling the exploration of new objects based on the current success rate in a naming game. Following this idea, we will study if NMF and LDA can support the definition of intrinsic motivation based on the current knowledge of objects and can provide improvement in the learning speed.

III. Proposed approaches

In order to compare NMF and LDA for object-word association learning, we use experimental data consisting of two channels: symbolic information for the language and continuous data for the visual perception that represents objects and the description of their shapes and colors. Noise and ambiguity exist in both channels since the visual presentation is sensitive to the changes in environmental conditions (e.g., lighting conditions) and may contain several objects, while the language description may contain words not related to the object identities or features (e.g., pronouns or errors from speech recognition). To highlight the importance of these factors, we designed two cases for the object ambiguity where either one or three objects are presented simultaneously to the system (named respectively "*single*" and "*triple*") and two cases for the language ambiguity: "*keywords only*" where the language channel has been manually corrected to contain only relevant words and "*full sentence*" where raw full sentences were used. Note that even in the "*single*","*keywords only*" scenario, ambiguity is present in the keyword-feature association as the two keywords may correspond to shape or color.

A. Data representation

As input for our models, we have a corpus V of vectors V_i $(i = 1, 2, ..., n)$ representing the appearance of an object and an associated sentence pronounced by a human partner (Figure 2). The first part of each vector is a continuous channel that represents features obtained through computer vision. These features are currently constructed to represent color (V_i^{color}) and shape (V_i^{shape}) of the object (see section IV-A), but they could be the results of a more generic feature computation algorithm. The features are encoded as vectors of constant size, and multiple objects of interest are represented by summing the description of each individual object, thanks to the fact that the features are histograms, which can be added. The second part of each vector is a binary vector of the size of the dictionary of all known words (V_i^{word}) and represents the word occurrences in the sentence. The dictionary is created incrementally, starting from an empty dictionary and adding each new word encountered in sentences at the end.

For the application of LDA and Term Frequency-Inverse Document Frequency (TF-IDF, see below), the non symbolic (visual) channel in the observation vectors in V needs to be quantized. The clustering is performed by a simple incremental clustering that puts each observation in the same cluster as a previous observation if its distance is smaller than a threshold (we used 0.7 in all subsequent experiments), or creates a new cluster otherwise. We use the χ^2 distance which is well adapted for histogram features :

$$\chi^2(x, y) = \sum_{k=1}^{d} (x_k - y_k)^2 / (x_k + y_k)$$

Each of the resulting shape cluster will be labelled as $s_t \in S$, while all member vectors within a cluster will be averaged as $v_{s_t} \in V_S$, then S and V_S act as entries and corresponding contents of the shape dictionary. The same procedure takes place for the formation of the color dictionary $\{C : V_C\}$. A corpus (D) of vector-quantized samples d_i, $(i = 1, 2, ..., n)$ is then established by finding the items $s_i \in S$ and $c_i \in C$ whose member vectors are most similar to V_i^{shape} and V_i^{color} respectively by applying χ^2 distance. Using the words $\mathbf{w_i}$ whose corresponding indices in V_i^{word} are positive, d_i indicates a collection of symbols, containing all words in $\mathbf{w_i}$ plus s_i and c_i.

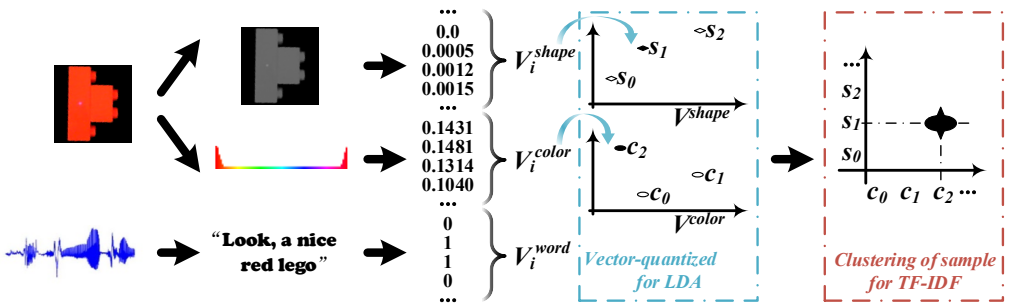

Fig. 2. Illustration of the data representation used in our experiments.

B. Language channel filtering

In the "full sentence" scenario, LDA will filter keywords thanks to its statistical properties, but NMF will provide better performance after an initial filtering of keywords [11].

The filtering method (see details in [11]) relies on statistics on the word occurrences through the Term Frequency-Inverse Document Frequency (TF-IDF) approach [19] popular in text processing. In the current paper, we computed the TF-IDF values using the clusters defined by the pairs (s_i,c_i) described in section III-A as documents. We improved the method by using an adaptive threshold on the Inverse Document Frequency value whose goal is to remove too common or too rare words. Thus we only retain words whose IDF value is between idf_{low} and idf_{high} defined as :

$$\begin{aligned} idf_{low} &= idf_{min} + \eta_{low}(idf_{max} - idf_{min}) \\ idf_{high} &= idf_{min} + \eta_{high}(idf_{max} - idf_{min}) \end{aligned} \quad (1)$$

where idf_{min} and idf_{max} are the maximum and minimum of idf values for all words. For the reported experiments, η_{low} and η_{high} values are optimized to reach the highest possible final performance in each scenario.

C. Learning using NMF

Using the V_i samples with the filtered linguistic part (or raw samples in the "keywords only" scenario), we use NMF [9] in order to discover reference vectors that explain data as sum of these vectors with positive weights. More precisely, NMF will find matrices W and H so that:

$$\begin{aligned} V_{m\times n} &\approx W_{m\times k} H_{k\times n} \\ \begin{bmatrix} V_{shape} \\ V_{color} \\ V_{word} \end{bmatrix}_{m\times n} &\approx \begin{bmatrix} W_{shape} \\ W_{color} \\ W_{word} \end{bmatrix}_{m\times k} [H_1, H_2, ..., H_n]_{k\times n} \end{aligned} \quad (2)$$

where V is the matrix containing the observations in columns, W, H are the matrices computed by NMF, W containing the k latent topics we are looking for and H being the weights to reconstruct the observations from the topics.

The W and H matrices are found by minimizing the following Kullback-Leibler divergence:

$$D_{KL}(V \,\|\, WH) = \sum_{ij} \left(V_{ij} ln \frac{V_{ij}}{(WH_i)_j} - V_{ij} + (WH_i)_j \right) \quad (3)$$

For this, we use the algorithm based on multiplicative updates proposed by Lee and Seung [20]. This method converges to a local minima, so the initialization is important. In order to favor the solution with one word for each reference element [11], we initialize the W_{word} matrix to the identity and (W_{shape},W_{color}) to random values.

D. Learning through LDA

LDA is used to infer statistical correlation between visual channel and keywords. Every sample d_i is thus seen as a collection of exchangeable discrete items ω_j (which can be colors c_i, shapes s_i or words w_i) and is modeled as a generative mixture model over a set of K hidden topics $\{z_1, \ldots, z_K\}$ defined by a probability distribution on the items $p(\omega_j, z_k, \beta)$. The likelihood of a sample is thus given by (see [10] for details):

$$L_{LDA}(d_i) = \int_\theta p(\theta|\alpha) \left(\prod_j \sum_{z_k} p(z_k|\theta) p(\omega_j|z_k, \beta) \right) \quad (4)$$

where $p(\theta|\alpha)$ is a Dirichlet distribution defining the topic mixture, $p(z_k|\theta)$ the probability of the topic z_k for this mixture and $p(\omega_j|z_k, \beta)$ the probability of the item for a given topic. The parameters $\Theta = \{\alpha, \beta\}$ of the model which have to be estimated includes the parameter α of the Dirichlet distribution, and the parameter β defining the probabilities $p(\omega_j|z_k, \beta)$ we are ooking for. Θ is estimated by maximizing the likelihood of the corpus

$$L_{LDA}(D) = \prod_{i=1}^{m} L_{LDA}(d_i)$$

using Collapsed Gibbs Sampling[1]. In practice, we observe that for a given k, the distribution $p(., z_k, \beta)$ is only significant for a couple (c_j, w_j) or a couple (s_j, w_j).

E. Incremental learning

In our tests, both NMF and LDA should deal with incremental learning, where new observations are added to the matrix V. While dedicated incremental learning algorithms

[1]We use the implementation from https://github.com/ariddell/lda with all parameters initialized with default settings

Fig. 3. The 39 objects used for the experiments.

could be used, in the current paper, we simply completely retrain the models using all the data of the updated matrix V and the corresponding D corpus. This approach therefore gives an upper bound of the performance an incremental learning algorithm could achieve.

F. Active learning

For active learning, we want to see if NMF and LDA can be used to define an efficient value function for the choice of the next samples. This value should estimate how well a sample is currently known, so we naturally use its reconstruction error estimated by the Kullback-Leibler divergence for NMF (eq. 3) and by its likelihood for LDA (eq. 4).

In the "single" scenario, we estimate $D_{KL}(V_i)$ or $L_{LDA}(d_i)$ for all the training samples and randomly select a sample among the 6 samples with the highest values. This slack strategy has been chosen to ensure some diversity in the object choice and was found important to improve the overall performances. For the "triple" scenario we select three objects among the ones with the top 12 values.

IV. Experimental results

A. Experimental setup

The experiment is conducted with a camera installed over a table, facing down to capture objects, and a microphone. 39 objects (Figure 3) are used, one at a time for recording, while a human teacher is describing it with complex sentences. The objects exist in 5 colors and 10 shapes, thus giving 15 keywords.

We used the image processing approach presented in [11] which segments the object, then produces a 900 elements shape descriptor, and a 80 elements color descriptor. The speech-text conversion uses Google speech-api[2] to convert a sentence into a word occurrence vector.

We recorded 153 samples with the help of ten volunteers, in which every object is described at least three times and most of them four times. Each object is described by two keywords, but the mean sentence length is 4.026, thus containing in average 2.026 irrelevant words. We create a training set by selecting 3 samples for each of the 39 objects (a total of 117 samples) and keep the remaining 36 samples, which cover all the keywords, as testing data to monitor the performance of learning.

[2]https://github.com/gillesdemey/google-speech-v2

For testing, we simulate the situation where the teacher utters a textual description encoded as a binary format T_j about an object j and the learner has to choose the right object from the pool of all 36 testing objects. For NMF, we first compute the coefficient vector of hidden topics H_i associated with the visual description of each testing object i by minimizing the distance $D_{KL}([V_i^{shape}, V_i^{color}] \parallel [W^{shape}, W^{color}]^T H_i)$, and reconstruct the textual description of each object: $V_i^{word} = W^{word} H_i$. We then find the object in the testing set whose textual description is the closest to T_j by computing $\chi^2(T_j, V_i^{word})$ for all i. We finally count the percentage of all right answers among the 36 testing objects.

For LDA, we estimate the hidden topic distribution associated to T_j: $P(\mathbf{z}|T_j)$, and reconstruct the associated vision feature channel using $P(\omega_j|T_j) = \sum_k P(\omega_j|z_k, T_j) \cdot P(z_k|T_j)$ where $\omega_j \in S \cup C$. Then for every testing sample d_i, we compute the log-likelihood $L(d_i|T_j) = \sum_l Cnt(\omega_l) \cdot \ln P(\omega_l|T_j)$, where $\omega_l \in S \cup C$ and $Cnt(\omega_l)$ is number of occurrence of visual cluster ω_l from the testing sample d_i. The object whose likelihood is the highest is taken as the answer and we compute the overall percentage of correct answers.

B. Experimental results

The proposed models are first tested for their overall learning abilities through a *batch learning* experiment given the complete set of training samples with keywords only (sec. IV-B1). We then evaluate their learning progress when training data are chosen incrementally from the set in a random manner, in the "keywords only" scenario (sec. IV-B2) and in the "full sentence" scenario in order to better approach realistic interactive scenarios (sec. IV-B3). We then demonstrate the effect of active learning strategy compared with random learning performance in the final experiment (sec. IV-B4).

*1) **Batch learning**:* In order to validate the overall performance and set the algorithm parameters (the number of topics), we use all "keywords only" data in the "single" and "triple" scenario. We were able to reach 100% performance in all cases. The ambiguity in multi-object cases ("triple") do not decrease performance; showing that both approaches are able to perform cross-situational learning efficiently with the number of samples considered.

For this, we set the number of topics in NMF to the number of total keywords (i.e., 15) as it achieves the best performance. For LDA, the best performance occurs when its topic number is higher (i.e., 20). However, the optimal number of topics was observed to depend on the data. Therefore, in the subsequent experiments, the number of topics is incrementally adapted by setting this number equal to the number of detected clusters from $S \cup C$ and then, at each learning step, letting it increase by 1 if the training with $n_{topics}+1$ produces larger overall log-likelihood than that with n_{topics}. This simple policy leads to optimal performances in our experiments despite the fact that the added topic do not correspond to keywords but account for a few noisy samples in the training data, representing either a feature description associated with a non-keyword symbol or a feature description without symbols.

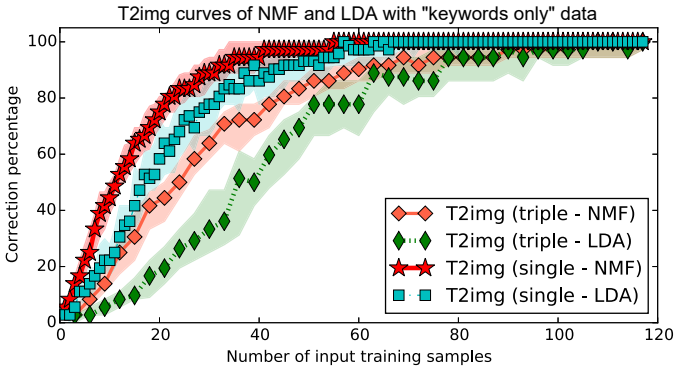

Fig. 4. Incremental learning with "keywords only" data.

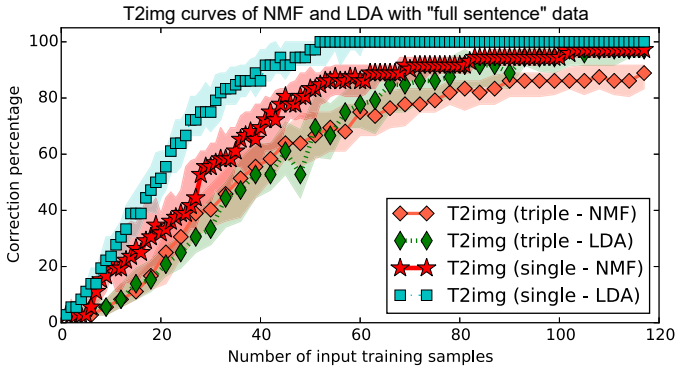

Fig. 5. Incremental learning with "full sentence" data.

*2) **Incremental learning with "keywords only" data**:* This experiment simulates the incremental learning scenario when a teacher randomly chooses an object (or a triple of objects) and describes it (or them) with their associated keywords. We report the testing performance as a function of the number of samples used for training, up to the total number of samples of 117. The curves display the 75_{th}, 50_{th} and 25_{th} percentile of performance among 50 repetitions of the experiments.

Figure 4 shows that despite a similar final performance of 100%, the learning progress of the methods appears different. First it is clear that learning in the "triple" case requires more samples than learning in the single case where a performance above 90% is reached already after 50 samples. We can also see that NMF consistently outperforms LDA regarding the learning speed in all cases, showing its adaptation to the case of limited ambiguities in the language part.

*3) **Incremental learning with "full sentence" data**:* This experiment is similar to the previous one, but using the full sentences. In figure 5, we observe that due to much more ambiguities in data compared to that of "keywords only", the performances are not guaranteed at the end of training to reach 100%, although this performance is still reached in the "single" case. Contrary to the previous scenario using only keywords, LDA learns much faster than NMF coupled with the statistical TF-IDF filtering and achieves higher final performances. This illustrates the better adaptation of the probabilistic model of LDA to this problem compared to NMF which requires a more complex pre-processing.

*4) **Active learning vs. random learning**:* The last experiment measures the performance of active learning compared to the random choice of samples. For this experiment, contrary to the previous ones, the training samples can be selected multiple times because all the 117 training samples are considered for the choice of the next sample using either the random or active strategy. This was made to highlight the ability of active learning to efficiently ignore the already known samples. Figure 6 shows the resulting performances for NMF and LDA.

We observe that active learning makes it possible to improve learning speed and performance in all scenarios, for both NMF and LDA, showing that a criterion relevant for active learning (section III-F) can be defined in both cases. However, the performance reached by NMF remains lower than those of LDA that achieves final almost perfect performances in all cases (see also table 6 below). We can also observe that the gain of active learning for LDA, compared to "single" scenario, is greater in the more ambiguous "triple" scenario, reaching full performance with only 60 samples.

TABLE I
AREA UNDER THE LEARNING CURVES OF FIGURE 6

	Active Learning		**Random Learning**	
	single case	*triple* case	*single* case	*triple* case
NMF	0.92	0.72	0.87	0.56
LDA	1	0.86	0.96	0.71

In order to quantify the differences between the cases, we computed the areas under the 50_{th} percentile learning curves from figure 6. Table I shows these values relative to the best performance obtained by LDA in the "single" case, with active learning. We observe that the overall worst approach is the use of NMF with random samples, and that the gain of using active learning with NMF is smaller than the gain of using LDA with random samples in "single" case, thus showing the importance of the learning approach over the learning strategy in our experiment.

V. DISCUSSION AND CONCLUSION

We compared two models of cross-situational learning of word meanings based on topic discovery algorithms, NMF and LDA. Both models achieved high performance in every experimental cases when there is a set of sufficient learning samples. They proved to be robust to both linguistic and referential ambiguities and both models were able to support active learning which was shown to accelerate the learning speed by comparison with random sample selection.

Each algorithm has its own better-suited scenario. NMF would be more adapted when dealing with only visual ambiguities and raw visual data ("keywords only" scenario), resulting in precise mono-modal concepts, once a correct number of components is provided. LDA shows better adaptability and robustness with clustered visual data when linguistic ambiguity and noise are involved ("full sentence" scenario) due to its

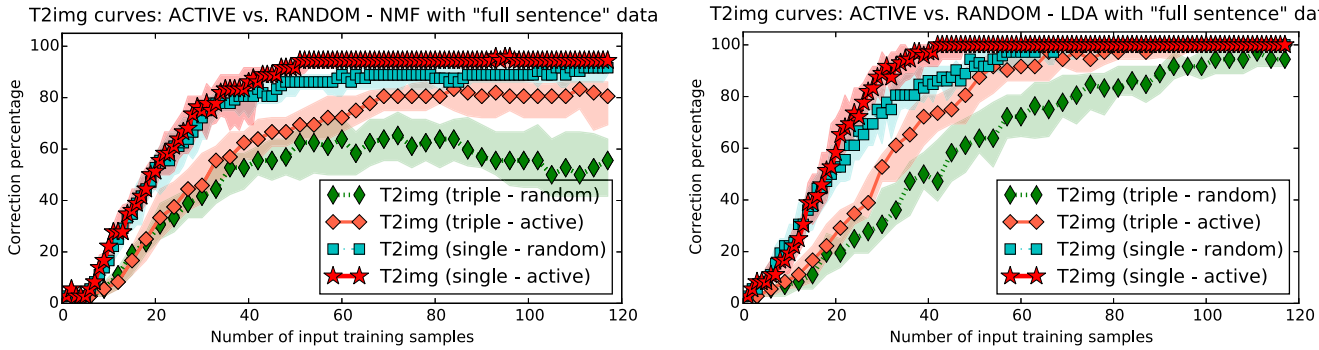

Fig. 6. Comparison between active learning and random learning by applying NMF (left) or LDA (right) with "full sentence" data

statistics-based nature. Contrary to this embedded mechanism of keywords selection in LDA, NMF has to be associated with a language filtering mechanism but is not able to reach similar performances in the "full sentence" scenario.

While our work is not intended at computational modeling of human performances, it is interesting to compare the active learning strategies implemented in our model to those used by humans. Kachergis et al. [15] shown that humans use various active strategies, but mainly rely on immediate sample repetition to facilitate learning. Yet from our implementation, the resulting strategy is different: random sample choices in the "triple" scenario led to a mean repetition of 2.42 words in successive steps, while the active choice led to a mean repetition of 1.89 words. Two basic reasons could be used to explain such a difference in applying the repetition strategy. On one hand, in [15], each trial consists of four mutually different objects thus no "within-trial repetition of objects" is allowed, however in our "triple" scenario experiment, the same features (shape or color) from different objects could appear in a triple and this gives rise to a "within-triple feature repetition" which can simply reduce the complexity of each triple. In fact, the number of repeated features inside a triple is 0.86 with the random strategy and 2.06 with the active choice. On the other hand, unlike computational models, humans are less efficient at keeping a long-term memory of the past co-occuring records and hence the successive repetition facilitates learning for humans but not for our model.

In future work, a better vision descriptor could be considered to record shape information, since the current pixel based method will obviously be limited in more realistic scenarios. We also plan to extend our approach to deal with homonyms, both for the language part and for the visual part, where an object can present different visual appearances depending on the observation point of view.

ACKNOWLEDGMENT

This work is supported by the China Scholarship Council.

REFERENCES

[1] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, pp. 335–346, 1990.

[2] L. Gleitman, "The Structural Sources of Verb Meanings," *Language Acquisition*, vol. 1, no. 1, pp. 3–55, 1990.

[3] M. Hirotani, M. Stets, T. Striano, and A. D. Friederici, "Joint attention helps infants learn new words: event-related potential evidence," *Neuroreport*, vol. 20, pp. 600–605, 2009.

[4] L. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," *Cognition*, vol. 106, no. 3, pp. 1558 – 1568, 2008.

[5] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive science*, vol. 26, pp. 113–146, 2002.

[6] C. Yu and D. H. Ballard, "A multimodal learning interface for grounding spoken language in sensory perceptions," *ACM Trans. Appl. Percept.*, vol. 1, no. 1, pp. 57–80, Jul. 2004.

[7] G. Kachergis and C. Yu, "Continuous measure of word learning supports associative model," in *Proc. of the IEEE International Conference on Development and Learning*, 2014, pp. 20–25.

[8] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[9] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing . . .* , pp. 556–562, 2001.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[11] Y. Chen and D. Filliat, "Cross-situational noun and adjective learning in an interactive scenario," in *Proc. of the 5th International Conference on Development and Learning (ICDL)*, Aug. 2015.

[12] T. Nakamura, T. Nagai, and N. Iwahashi, "Grounding of word meanings in multimodal concepts using lda," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 3943–3948.

[13] P. Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, April 2007.

[14] W. Schueller and P.-Y. Oudeyer, "Active Learning Strategies and Active Control of Complexity Growth in Naming Games," in *Proc. of the 5th International Conference on Development and Learning (ICDL)*, 2015.

[15] G. Kachergis, C. Yu, and R. M. Shiffrin, "Actively learning object names across ambiguous situations," *topiCS*, vol. 5, no. 1, pp. 200–213, 2013.

[16] O. Mangin, D. Filliat, L. ten Bosch, and P.-Y. Oudeyer, "MCA-NMF: Multimodal Concept Acquisition with Non-Negative Matrix Factorization," *PLoS ONE*, vol. 10, no. 10, p. e0140732, Oct. 2015.

[17] T. Nakamura, Y. Ando, T. Nagai, and M. Kaneko, "Concept formation by robots using an infinite mixture of models," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015, pp. 4593–4599.

[18] F. Kaplan, P.-Y. Oudeyer, and B. Bergen, "Computational models in the debate over language learnability," *Infant and Child Development*, vol. 17, no. 1, pp. 55–80, 2008.

[19] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Aug. 1988.

[20] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization." *Nature*, vol. 401, pp. 788–791, 1999.